Teaching and Assessing Statistical Approaches to Machine Translation

May 30, 2003

Abstract

Empirical methods in Natural Language Processing (NLP) and Machine Translation (MT) have become mainstream in the research field. Accordingly, it is important that the tools and techniques in these paradigms be taught to potential future researchers and developers in University courses. While many dedicated courses on Statistical NLP can be found, there are few, if any courses on Statistical Approaches to MT. This paper presents the development and assessment of one such course as taught to final year undergraduates taking a degree in NLP.

1 Introduction

It is relatively uncontroversial to state that empirical methods in Natural Language Processing (NLP) and Machine Translation (MT) have become mainstream in the research field. Accordingly, it is important that the tools and techniques in these paradigms be taught to potential future researchers and developers in University courses. Many degree courses nowadays contain specific modules on statistical NLP (as it relates to word-sense disambiguation, parsing, generation etc.). While more and more courses on MT address the statistical approaches which are currently in vogue, it is more a case that this is done in passing in a couple of classes, rather than devoting a whole module to statistical approaches to MT. Indeed, in a trawl of the Web, we could find no such specific courses on statistical paradigms in MT. Of course, that is not to say that none exist; merely that none could be found by us.

Naturally, courses on MT need to take into account the students' skills and demands. Kenny & Way (2001) describe how MT is taught in one institution to students of differing backgrounds. A distinction is made in that paper between *users* versus *developers*: while language students and translators can be expected to be able to use translation tools in their careers as translators, students of NLP with a specialisation in MT might realistically be employed as designers and implementors of such tools in a programming or localisation environment.

This paper presents a course on statistical methods in MT taught to final year undergraduates taking a degree in NLP, focussing mostly on Example-Based MT (EBMT). These students have a strong background in programming, language skills and good competence levels in formal linguistics and NLP. Accordingly, the course is very practically oriented, and the students are expected by the end of the course to be in a position to develop a toy EBMT system. The presentation of the course in this paper is not intended to be prescriptive; rather, in reporting on the chosen methodology, it may serve as a basic model for others considering teaching such a module to similar students. In addition, by providing details of what worked and—more importantly, what did not—the hope is that others may benefit from lessons learnt.

The rest of the paper is organised as follows: in section 2, we discuss why teaching statistical approaches to MT is becoming ever more important, and provide some documentation of where such material is covered. Section 3 describes the content of the course taught be these authors. Section 4 describes the method of assessing the students to whom this course was delivered, and reports on the lessons learned. Finally we conclude, and provide possible improvements to the course.

2 Teaching Statistical Approaches to MT

It is fair to say that statistical approaches to NLP and MT have matured to the extent that they may now be considered mainstream. Indeed, at any major conference on NLP/MT, one can expect to encounter more papers which favour an empirical approach than those which utilise rules and/or constraints.

If University students are to become familiar with such techniques and tools, dedicated modules have to be designed which address these topics. However, while there are many courses on statistical NLP already in existence, in a trawl of the Web, we could find no courses which solely address statistical approaches to MT. A similar parallel can be drawn with respect to textbooks: there are a number of volumes which specifically address empirical methods in NLP (e.g. Charniak, 1993; Manning & Schütze, 1999), but no textbooks are geared solely towards the concerns of statistical methods in MT. This hole in the market will soon be partially filled by (Carl & Way, 2003), which in turn may see an increase in the number of courses on EBMT, but there is still a need for a book on Statistical MT (SMT) itself (e.g. Brown *et al.*, 1990, 1992; Yamada & Knight, 2001; Soricut *et al.*, 2002).

Nevertheless, while there would appear to be no courses (other than the one described in the next section) dedicated solely to the teaching of statistical methods in MT, SMT and EBMT have become so well established that any contemporary course on MT would be incomplete without at least equipping students with some superficial knowledge of these techniques. Some examples of courses which address these newer statistical approaches include, but are not limited to, the following:

- MSc. in MT, CALL and NLP at UMIST, UK;[1]
- MSc./Ph.D. Program in Language and Information Technologies at CMU, Pittsburgh.[2]

Others address the topic in modules on Empirical NLP, including:

- Programmes in Computer Science, ISI, CA;[3]
- Undergraduate Study in Computer Science at Brown University, Providence, RI;[4]
- Postgraduate programmes in Computer Science at UMIACS, MD.[5]

In addition, some of the newer textbooks on NLP/MT address these and related issues, e.g. Trujillo (1999, Chapter 8) goes into some detail on EBMT and SMT; Bowker (2002, Chapter 5) discusses the related area of Translation Memory (TM) systems; and to a lesser extent, Jurafsky & Martin (2001, Chapter 21) provide some detail on how statistical techniques can be used in MT. Nevertheless, until the advent of recent books dedicated to statistical methods in MT (e.g. Carl & Way, 2003), instructors in this area have had to rely on original papers and survey articles (e.g. Somers, 1999).

3 Course Content

Kenny & Way (2001) contrasts how courses on MT have to be tailored towards different sets of students with different backgrounds, even in the same institution. One of the authors of this paper also teaches a basic introduction to EBMT in two hours to a group of postgraduate students taking a degree in Translation Studies (TS). This is a small component of a module on Translation Technology. Of course, given that TS students students are more interested in TM tools, a superficial overview of EBMT, and especially the differences between TM and EBMT, suffices for this group of students. While both EBMT and TM require

[1] http://www.ccl.umist.ac.uk/teaching/modules/3000/3003.htm,http://www.ccl.umist.ac.uk/teaching/modules/5000/5092.html
[2] http://www.lti.cs.cmu.edu/Courses/11-731-desc.html
[3] http://www.isi.edu/natural-language/people/cs562-2003.htm
[4] http://www.cs.brown.edu/courses/cs241/
[5] http://benreilly.umiacs.umd.edu/~hwa/cmsc828-02/

aligned corpora, for instance, the TS students are far more likely to *use* built-in alignment tools such as Trados *WinAlign*, NLP students may be expected to *develop* their own alignment software.

The course presented here is geared specifically to a group of final year undergraduate students taking a degree in NLP. It would, therefore, be an inappropriate model for groups of students with differing backgrounds. The course in Statistical Approaches to MT taught by us consists of 3 hours a week lectures and a 2 hour practical session over a period of 8 weeks. The content of the course is as follows:

- Week 1:
  - Double lecture: Revision class on Perl.
  - Single lecture: Introduction to Course and Statistics-based MT.
- Week 2:
  - Lab: Perl exercises.
  - Double lecture: Corpus-based language and translation models. What corpora to use, and how to get them.
  - Single lecture: Alignment Methods: word-based, character-based etc.
- Week 3:
  - Lab: Sentential-level alignment using:
    - relative sentence position;
    - relative length of sentence.
  - Double lecture: Probabilistic Word (and Phrase) Models.
  - Single lecture: How to improve alignment with (simple set of) heuristics (see Lab, Week 4).
- Week 4:
  - Lab: Add in cognates, paragraph markers, punctuation, HTML tags and other anchors to improve alignment.
  - Double lecture: EBMT. How it works, comparison with TM etc.
  - Single lecture: Marker Hypothesis as segmentation tool.
- Week 5:
  - Lab: Build word-alignment tool (i.e. probabilistic lexicon) using mutual information
  - Double lecture: Marker Hypothesis: advantages/disadvantages.
  - Single lecture: Problems for EBMT:
    - boundary definition;
    - boundary friction;
    - what examples to store etc.
- Week 6:
  - Lab: Build and Test EBMT system using 500 sentences of English, French and German data. Use sententially- and word-aligned databases.
  - Double lecture: Generalised Templates (cf. rules in rule-based methods in MT).
  - Single lecture: Generalised Templates.

- Week 7:
  - Lab: Improve model with generalised templates and word insertion.
  - Double lecture: Examples of EBMT systems.
  - Single lecture: Towards Hybrid Models.
- Week 8:
  - Lab: Finalise EBMT System.
  - Double lecture: Statistical MT.
  - Single lecture: Statistical MT.

The basic model followed here was that the material delivered in class during one week was put into practice in the labs during the following week. Essentially, the course is split into three chunks, namely Alignment (both sentential and word-level), EBMT and SMT.

In the introductory part of the course, students are made aware of the need for statistical language and translation models to be developed from large, good quality, representative monolingual and bilingual corpora. By concocting toy corpora which do not fulfill these criteria, and asking students to calculate a number of unigram and bigram probabilities based on data contained in these corpora, it is quite easy to demonstrate that a number of undesirable effects follow when small, unrepresentative corpora are used. The advantages and disadvantages of bigram models are then presented to the students.

Despite the fact that the various mathematical techniques employed are, in principle at any rate, utilisable for any pair of languages, the fact that sententially aligned bilingual corpora exist only for a few language pairs renders these techniques somewhat less generally applicable. In order to try to overcome this problem, some consideration is given to using the Web as a corpus from which usable bitexts might be extracted (cf. Grefenstette, 1999; Resnik & Smith, 2003).

Some of the major algorithms for aligning bilingual corpora are then presented (Brown *et al.*, 1991; Gale & Church, 1993; Kay & Röscheisen, 1993). These are interestingly different, in that the method of Brown *et al.* uses a length-based metric which counts words, that of Gale & Church uses a character-based model, while Kay & Röscheisen require the use of a bilingual dictionary.

As for sub-sentential alignments, students are shown how to estimate co-occurrence using Mutual Information. With particular respect to EBMT, other methods of segmentation are also presented, especially Marker-Based segmentation (e.g. Veale & Way, 1997; Way & Gough, 2003). Students are also made aware of the need for extracted sub-sentential alignments to be made more general, in order to improve coverage and robustness. Some of the techniques presented include:

- Extracting transfer rules from examples (e.g. Furuse & Iida, 1992);
- Generalising by syntactic category (e.g. Kaji *et al.*, 1992);
- Generalising by semantic features (e.g. Matsumoto & Kitamura, 1995);
- Generalising Parse Trees (e.g. Way, 2003);
- Generalising Strings (e.g. Cicekli & Güvenir, 2003; McTait, 2003);
- Generalising using Placeables (e.g. Brown, 1999).

These methods are embedded in a basic outline of the EBMT process. Comparisons are made with TM, some experience of which the students have had before. The problems of boundary definition and boundary friction are presented, storage of examples is discussed, issues pertaining to segmentation are put forward, and the matching and recombination stages of EBMT are explained. We then present the IBM Models 1 and 2 of SMT. Finally, given that these students have previously taken a module on RBMT, we discuss possible hybrid models which combine elements of both paradigms as a more effective solution to the problems of translation.

4 Assessment

The course was designed to be an 'assessment only' module (i.e. no end of module exam), for two reasons. Firstly, being a second semester course, any final examinations would be scheduled during time which students would otherwise be spending on implementing their final year project (which constitutes 33% of their overall degree classification). Secondly, students at our University are classified only on their final year marks, so if students were to fail either the exam or assessment component of any final year module, they would not be eligible for an Honours degree. It was felt, therefore, that by having the module evaluated purely by continuous assessment, these issues could be best avoided.

There were two assignments:

1. a labtest on building an aligner (week 5);
2. a group presentation/demonstration on building an EBMT system (week 8).

The labtest was a 3-hour assessment, in which the students were individually asked to develop a number of programs in Perl, namely:

- to calculate the average sentence length of the ⟨*source,target*⟩ sentences provided in terms of both words and characters;
- to calculate the ratio of ⟨*source,target*⟩ words and characters per sentence;
- to write a length-based sentence aligner, in terms of both words and characters;
- to compare the alignment results against a 'gold standard' provided;
- to segment the 'gold standard' reference solution according to the marker hypothesis;
- to propose sub-sentential alignments using the marker hypothesis.

In addition, there were three discussion questions on aspects of the course.

Note that all of these programs had been tackled in the lab sessions during the course. We considered three hours to be a reasonable time limit given that one of the authors was able to write programs to perform the various tasks in one hour. Nevertheless, we found that we had overestimated quite considerably the amount we thought the students were capable of in the time available: none of them completed all questions, and in general, too many students spent far too long on programs for which very few marks were awarded (as indicated on the question paper).

While the students' answers were marked benevolently (for instance, where students provided pseudocode instead of actual Perl code, full marks were given if the pseudocode was a complete solution to the problem at hand), using the original schema, over half of the class had failed, with the top mark being just 57%. However, a compromise was developed whereby the marks were divided by 0.7 in order to give a truer indication of each student's performance (top mark 81%, lowest 19%, average 53%).

The second assessment was a group project, where the students were divided into groups of three and were asked to develop an EBMT system, based on the sentence-, phrasal- and word-level alignments written in preparation for the first assignment. Marks were awarded both for system design and functionality, and for documentation. No one segmentation method was preferred over any other; indeed, some groups used the marker-based approach, others used a bigram approach, etc. The groups presented their systems to these authors, who found their efforts to be extremely good (highest mark 90%, lowest 57%).[6] Finally, in order to derive the final mark for the module, the first assignment was weighted 0.35, with the group assignment weighted by 0.65. All students passed the module (highest mark 87%, lowest 44%, average 65%).

[6]For an example, consult `http://www.computing.dcu.ie/~sfoy-cl4/ebmt.html`

5 Conclusion

Statistical approaches to NLP and MT have reached a reasonable stage of maturity. It is important, therefore, that the tools and techniques underpinning these fields be taught to University students, who are likely to form a pool from which future researchers and developers in these areas are to be found. While dedicated courses on statistical NLP have made their way into many University curricula, we were unable to find any courses on empirical methods in MT in a trawl of the Web. Similarly, while a number of textbooks have appeared on statistical NLP, the first such book on one of the flavours of statistical MT, namely EBMT, is only just about to appear.

This paper presents the development and assessment of one such course as taught to final year undergraduates taking a degree in NLP. It focusses mainly on Alignment, EBMT and SMT. It was designed so that student performance was evaluated purely in terms of continuous assessment. We commented on the problems that arose in scheduling a laboratory test of the students' understanding of alignment, both at the sentential and sub-sentential levels. Nonetheless, when asked to develop an EBMT system in small groups, the students rose to the occasion.

As for improvements/changes to the course, the students may be assessed on a more ongoing basis in weekly labs rather than in one laboratory examination. In addition, the material in (Carl & Way, 2003) may be used in study classes, with students presenting this material to the class on a weekly basis. Finally, we may choose to focus more on the building of SMT systems, using the excellent, freely available Egypt toolkit.[7]

In sum, despite some teething problems, it can be said with some confidence that the module was successful. The developers and teachers of this course have certainly learned from the experience, and it is hoped that others who are considering the development of similar courses may find some value in the sharing of our experiences. That said, nothing in this paper is intended to be prescriptive: if the course structure and methods of assessment are of use to others, then fine, but if some other model is chosen, then we too would hope to benefit from the development of similar or related material.

References

[1] Bowker, L. (2002): *Computer-Aided Translation Technology*, University of Ottawa Press, Ottawa, Canada.

[2] Brown, P., J. Cocke, S. Della Pietra, F. Jelinek, V. Della Pietra, J. Lafferty, R. Mercer and P. Rossin (1990): 'A Statistical Approach to Machine Translation', *Computational Linguistics* **16**:79–85.

[3] Brown, P., J. Lai and R. Mercer (1991): 'Aligning sentences in parallel corpora', in *Proceedings of 29th Annual Meeting of the Association for Computational Linguistics*, University of California, Berkeley, pp.169–176.

[4] Brown, P., S. Della Pietra, V. Della Pietra, J. Lafferty and R. Mercer (1992): 'Analysis, Statistical Transfer, and Synthesis in Machine Translation', in *4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal, Canada, pp.83–100.

[5] Brown, R. (1999): 'Adding Linguistic Knowledge to a Lexical Example-based Translation System', in *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 99)*, Chester, England, pp.22–32.

[6] Carl, M. and A. Way (eds.) (2003): *Recent Advances in Example-Based Machine Translation*, Kluwer Academic Publishers, Dordrecht, The Netherlands (in press).

[7] Charniak, E. (1993): *Statistical Language Learning*, MIT Press, Cambridge, MA.

[7] `http://www.clsp.jhu.edu/ws99/projects/mt/`

[8] Cicekli, I. & H.A. Güvenir (2003): 'Learning Translation Templates from Bilingual Translation Examples', in M. Carl & A. Way (eds.) *Recent Advances in Example-Based Machine Translation*, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp.255–286.

[9] Furuse, O. and H. Iida (1992): 'An Example-Based Method for Transfer-Driven Machine Translation', in *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation: Empiricist vs. Rationalist Methods in MT (TMI-92)*, Montréal, Canada, pp.139–150.

[10] Gale, W.A. and K.W. Church (1993): 'A Program for Aligning Sentences in Bilingual Corpora', *Computational Linguistics* **19**(1):75–102.

[11] Grefenstette, G. (1999): 'The World Wide Web as a Resource for Example-Based Machine Translation Tasks', in *Proceedings of the ASLIB Conference on Translating and the Computer* **21**, London, [pages not numbered].

[12] Jurafsky, D. and J. Martin (2000): *Speech and Language Processing*, Prentice Hall, Upper Saddle River, NJ.

[13] Kaji, H., Y. Kida and Y. Morimoto (1992): 'Learning Translation Templates from Bilingual Text', in *COLING-90, Papers Presented to the 13th International Conference on Computational Linguistics*, Helsinki, Finland, pp.672–678.

[14] Kay, M. & M. Röscheisen (1993): 'Text-translation alignment', *Computational Linguistics* **19**(1):121–142.

[15] Kenny, D. and A. Way (2001): 'Teaching Machine Translation & Translation Technology: A Contrastive Study', in M. Forcada, J-A. Pérez-Ortiz & D. Lewis (eds) *Proceedings of the Workshop on Teaching Machine Translation, MT Summit VIII*, Santiago de Compostela, Spain, pp.13–17.

[16] Manning, C. and H. Schütze (1999): *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA.

[17] Matsumoto, Y. and M. Kitamura (1995): 'Acquisition of Translation Rules from Parallel Corpora', in R. Mitkov and N. Nicolov (eds.) *Recent Advances in Natural Language Processing: Selected Papers from the Conference*, John Benjamins, Amsterdam, pp.405–416.

[18] McTait, K. (2003): 'Translation Patterns, Linguistic Knowledge and Complexity in EBMT', in M. Carl & A. Way (eds.) *Recent Advances in Example-Based Machine Translation*, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp.307–338.

[19] Resnik, P. and N. Smith (2003): 'The Web as a Parallel Corpus', *Computational Linguistics* **29**(3) (to appear).

[20] Somers, H. (1999): 'Review Article: Example-based Machine Translation', *Machine Translation* **14**(2):113–157 (revised and updated in M. Carl & A. Way (eds.) (2003) *Recent Advances in Example-Based Machine Translation*, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp.3–57).

[21] Soricut, R., K. Knight and D. Marcu (2002): 'Using a Large Monolingual Corpus to Improve Translation Accuracy', in S. Richardson (ed.) *Machine Translation: From Research to Real Users, 5th Conference of the Association for Machine Translation in the Americas (AMTA-2002)*, LNAI 2499, Springer Verlag, Berlin/Heidelberg, Germany, pp.155–164.

[22] Trujillo, A. (1999): *Translation Engines: Techniques for Machine Translation*, Springer, London.

[23] Veale, T. and A. Way (1997): '*Gaijin*: A Bootstrapping, Template-Driven Approach to Example-Based Machine Translation', in *International Conference, Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria, pp.239–244.

[24] Way, A. (2003): 'Translating with Examples: The LFG-DOT Models of Translation', in M. Carl & A. Way (eds.) *Recent Advances in Example-Based Machine Translation*, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp.443–472.

[25] Way, A. and N. Gough (2003): '*wEBMT*: Developing and Validating an Example-Based Machine Translation System using the World Wide Web', *Computational Linguistics* **29**(3) (to appear).

[26] Yamada, K. and K. Knight (2001): 'A Syntax-Based Statistical Translation Model, in *39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics*, Toulouse, France, pp.523–530.